REMARKS

This Amendment is submitted in response to the Office Action dated February 22, 2007, 2006, having a shortened statutory period set to expire May 22, 2007. The present amendment amends Claims 15, 21 and 24. Upon entry of the proposed claims, Claims 15-17 and 21-24 will now be pending.

Rejections 35 U.S.C. § 103

In paragraph 4 of the present Office Action, Claims 15-17 and 21-24 are rejected as being unpatentable over *Mantha et al.* (U.S. Patent No. 6,163,779 – "*Mantha*") in view of *Wyler* (U.S. Patent No. 7,047,033 – "*Wyler*"). Applicants respectfully traverse these rejections.

With regards to exemplary **Claim 15**, a combination of the cited art does not teach or suggest "identifying unnecessary information elements in the HTML document, wherein the unnecessary information elements include...a block of text in the HTML document that is shorter than a maximum predetermined length, and wherein the block of text appears in the HTML document more than a predetermined frequency," as supported at paragraph [0085] of the current disclosure. That is, unnecessary information includes multiple repetitions of a same short block of text. The Examiner cites *Wyler* on col. 31, line 65 to col. 32, line 7, and col. 32, lines 47-52, which state:

> The "logical location" of an object which is interiorly disposed relative to the base object is the maximum value e.g. 100. The "logical location" of any other object is the distance, on the webpage, of that object from the base object."
>
> "Classifying one or more objects as cardinal: As described, a base object is selected which is the largest object on the webpage. If there is a tie, i.e. if the largest two or more objects are similar, to a predetermined extent, in size, then the object with the most words in it is typically deemed to be the base object."

Thus, these passages teach that the location of an object can be described as a distance to the largest object on a webpage. Applicants respectfully traverse the Examiner's position that this is equivalent to multiple short blocks of text being located on an HTML document too many times.

Furthermore, a combination of the cited art does not teach or suggest "a block is deemed to be meaningless if that block contains only said unnecessary information elements and at least one anchor," as supported in paragraph [0085] of the present specification. The Examiner cites *Wyler* on col. 12, lines 4-8, which states:

> In this level the application removes irrelevant information (images and data i.e. advertising banners, links to unrelated issues) from the webpage, and reorganizes the information into objects with categories in a file represent by the M2O script language.

That is, *Wyler* teaches that images such as advertising banners and links to other irrelevant pages can be purged from a webpage. However, there is no teaching or suggestion of "multiple short block of text being repeated" (as discussed above), "multiple anchors having a same title," "image tags that only perform a role of punctuation," and "text block having a same description" as being requisite components for defining a meaningless block in an HTML document. Similarly, there is no teaching or suggestion that this meaningless block in the HTML document has "at least one anchor."

Furthermore, a combination of the cited art does not teach or suggest "crawling only anchors found in blocks that have not been defined as OBJECT_DELIMETERS" (i.e., only crawling anchors that are not meaningless). The Examiner cites col. 14, lines 39-44 of *Wyler* for this teaching. This passage states:

> Second, the application searches all the documents for words that fit into the Index category, and when finding such words, the application inserts an "index" command. From that point on, the webpages are called "documents" since they have no longer have properties of a webpage.

This passage states that a crawler ("application") searches all documents for key words ("that fit into the Index category"). There is no teaching or suggestion of crawling only anchors that are not composed exclusively of "plural information elements that include an

OBJECT_IMAGE having a same Uniform Resource Locator (URL), wherein the OBJECT_IMAGE describes a type of media used to display the HTML document, a block of text in the HTML document that is shorter than a maximum predetermined length, and wherein the block of text appears in the HTML document more than a predetermined frequency, multiple anchors having a same title, image tags that only perform a role of punctuation for text in the HTML document, and multiple text blocks having a same description."

Therefore, in light of the present amendment further distinguishing the definition of "unnecessary information," Applicants respectfully request that the rejection of **Claims 15, 21 and 24** be withdrawn.

With regards to exemplary **Claim 16**, a combination of the cited art does not teach or suggest "the maximum predetermined length (of the block of text) is 12 bytes." For this feature, the Examiner cites *Wyler* at col. 11, lines 27-28; col. 13, lines 58-61; and col. 32, line 50, which state:

> Base - The object which is the biggest or has the most Object number of words in it.
>
> Object size - the object size is a value equal to Width * Height of the object
>
> (T)he object with the most words in it is typically deemed to be the base object

The cited passages state that the biggest size of a base object (see above for discussion of what a "base object" denotes) may be determined. Applicants respectfully traverse the Examiner's statement that this is equivalent to a block of text having a maximum size (12 bytes), which is used to denote the block of text as being "unnecessary."

Applicants therefore request that the rejection of **Claims 16 and 22** be withdrawn.

Similarly, with regards to exemplary **Claim 17**, a combination of the cited art does not teach or suggest that "the predetermined frequency (of occurrences of the block of text) is ten times." For this feature, the Examiner cites *Wyler* on col. 16, lines 39-41; col. 19, lines 3-7 and 33-36; and col. 31, line 50, which state:

> The mechanism of selecting the relevant objects is based on selecting the objects with weights that pass the predefined thresholds. In FIG. 4 we can see (marked by diagonal lines) a relevant region that passes the predefined thresholds.
>
> In this phase, the application tries to reduce the Index list length by finding identical words with different page numbers. The application then indicates the word followed by a list of all the reference page numbers.
>
> If Chapters and Sections with Titles and Sub-Titles do not appear in the document after the third level, only the following changes typically take place: 1. In the second level--irrelevant images/data are taken off.
>
> Typically, the "word matching" property is computed by performing a key word matching process. In this process, each word within the object whose "word matching" property is being computed, is taken up in turn and the system determines whether this word occurs in the base object. The system counts the number of words in the object which do occur in the base object. The proportion of words in the object which occur in the base object, from among the total number of words in the object, typically determines the "word matching" property of the object.

Applicants understand the Examiner's position to be, in essence, that a passage can be determined as being significant to a crawler if a particular word is found multiple times in a base (large) object, and that this operation is equivalent to a block of text occurring more than ten times causing a block of an HTML document to be deemed meaningless. Applicants respectfully disagree, since the two concepts are diametrically opposed. That is, *Wyler* teaches that any passage having multiple entries (of a word on a document) is <u>meaningful</u> (as per any standard crawling technique). Conversely, the present invention states that multiple entries (of a block of data in a document) make that block of data <u>meaningless</u>.

Applicants therefore request that the rejection of **Claims 17 and 23** be withdrawn.

CONCLUSION

As the cited art does not teach or suggest all of the presently claimed features, Applicants now respectfully request a Notice of Allowance for all pending claims.

No extension of time for this response is believed to be necessary. However, in the event an extension of time is required, that extension of time is hereby requested. Please charge any fee associated with an extension of time as well as any other fee necessary to further the prosecution of this application to **IBM CORPORATION DEPOSIT ACCOUNT No. 09-0461.**

Respectfully submitted,

James E. Boice
Registration No. 44,545
DILLON & YUDELL LLP
8911 North Capital of Texas Highway
Suite 2110
Austin, Texas 78759
512.343.6116

ATTORNEY FOR APPLICANT(S)